Project Performance

Jessica Zhang

April 26, 2018

Approach

- Irrelevant data could cause unnecessary noise and performance drop. So before analysis, we will need to remove irrelevant data.
- Null data needs to be dealt with as well.

- Naive analysis will be performed on the dataset as the baseline.
- Linear Regression model will be used for initial analysis.
- Random Forest and Expectation Maximization will also be used for analysis.

Data cleaning

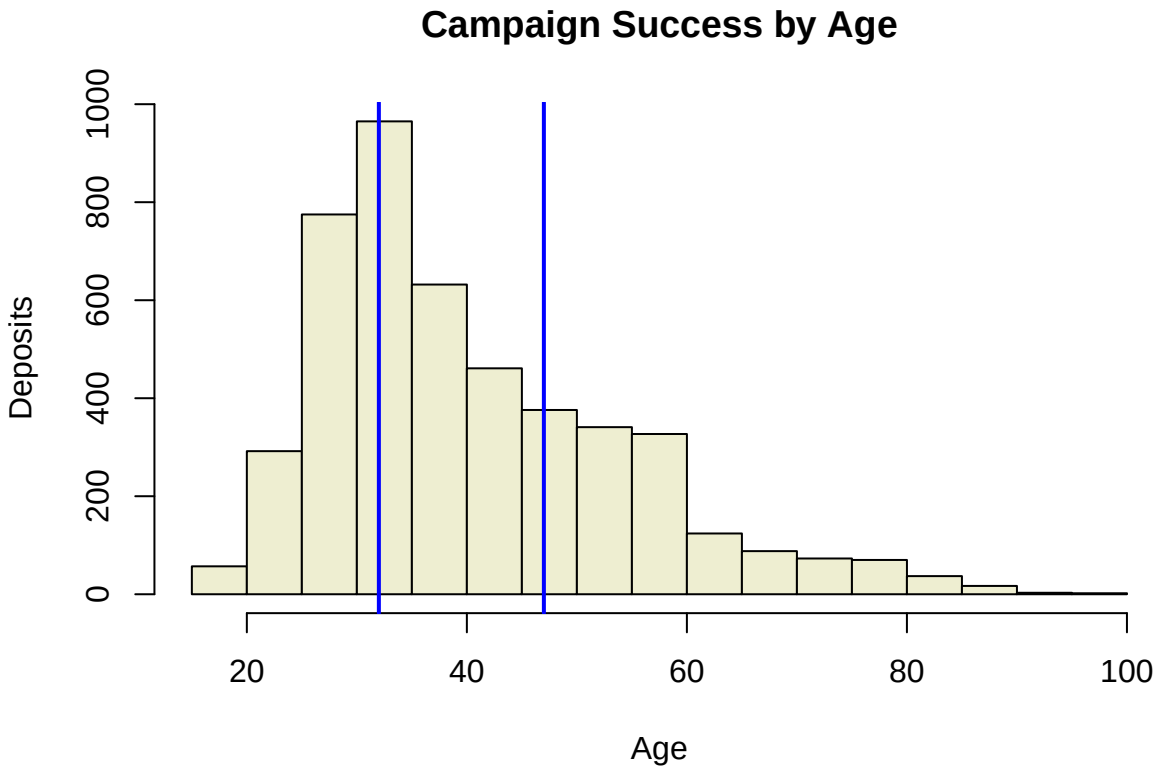
Columns 9-20 contains the campaign related information and are not related to clients' profile. So we have removed the extra information and saved the new data to another file named "updatedbankingdata.csv." In the rest of this project, we will be using data from updatedbankingdata.csv.

The updated data has totally 9 columns of data:

1. age (numeric)
2. job : type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid","management","retired","self-employed","services","student","technician","unemployed","unknown")
3. marital : marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed
4. education (categorical: "basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree"
5. default: has credit in default? (categorical: "no","yes","unknown")
6. housing: has housing loan? (categorical: "no","yes","unknown")
7. loan: has personal loan? (categorical: "no","yes","unknown")
8. contact: contact communication type (categorical: "cellular","telephone")
9. y: was the campaign successful? ("no", "yes")

Naive Analysis – The Age Group Assumption

We assume the effectiveness of the campaign is somehow related to the age of the clients and we look at the number of success each age group generated in the past campaign, we can see the deposit distribution in the figure below:



As we can see from the graph, the middle of the age group yields most deposits. A simple solution will be for the campaign team to focus on the middle 50 percentile of the client base and call the remaining later. The deposits will be generated during the busy season would be: 1975

The total number of deposites generated if half the clients were randomly called would be: 2320

This approach is less effective than the randomly picked clients. Therefore, this age group model is not a good model.

Simple Analysis – Linear Regression

Next, we use the linear regression to fit the data. Below is a summary of the model.

	FALSE	TRUE
NO	10622	11
YES	3281	11

Training Summary

Call: glm(formula = y ~ job + marital + education + loan + contact, family = binomial, data = train)

Coefficients:

```
                Intercept |                jobblue-collar
```

```
                  -1.07483 |                     -0.32906
           jobentrepreneur |                  jobhousemaid
                  -0.98544 |                     -0.80207
             jobmanagement |                   jobretired
                  -0.47488 |                      0.20876
         jobself-employed |                   jobservices
                  -0.88812 |                     -0.60546
                jobstudent |                 jobtechnician
                   0.33587 |                     -0.24126
             jobunemployed |                maritalmarried
                  -0.56344 |                      0.45415
             maritalsingle |             educationbasic.6y
                   0.38717 |                     -0.21560
         educationbasic.9y |          educationhigh.school
                  -0.15356 |                      0.11057
       educationilliterate |  educationprofessional.course
                   0.14821 |                      0.02112
educationuniversity.degree |                         loan
                   0.35364 |                     -0.55181
          contacttelephone |                     -0.93525
```

Degrees of Freedom: 11120 Total (i.e. Null); 11100 Residual
Null Deviance: 11220
Residual Deviance: 10520 AIC: 10560

Confusion Matrix:

	FALSE	TRUE
NO	8842	11
YES	2257	11

Observations:

1. The deviance residuals are not symmetrical, which indicates the model may not fit the data well.
2. The coefficients matrix shows that some of the parameters are more related to the results than others. Parameters seem to be able to influence campaign output are: Job, Education, Marital, Loan, and Contact Parameters seem to be irrelevant to the output are: Age and Default.
3. The large number of deviance and degrees of freedom further indicates that the model is not a good fit for this dataset.
4. The confusion matrix shows that even though most "no" lables are predicted correctly, most "yes" lables are mistakenly predicted as "no." This model adds no apparent value to our goal of improving the marketing effectiveness.

Further Analysis – Random Forest

Call: randomForest(formula = y ~ ., data = mydata, importance = TRUE) Type of random forest: classification Number of trees: 500 No. of variables tried at each split: 4

```
    OOB estimate of  error rate: 8.52%
```

Confusion matrix: no yes class.error no 35254 1294 0.03540549 yes 2215 2425 0.47737069

	FALSE	TRUE	Class.error
NO	35254	1294	0.03540549
YES	2215	2425	0.47737069

Observations:

1. The accuracy of the model is good. We can predict 80 percent of the yes responses.

2. However, more than 90 percent of positive response was predicted as negative. The result is better than linear regression, but not good enough for improving the marketing effort successfulness.

Further Analysis – Expectation Maximization